2018 기술경영경제학회 하계 학술대회

재무 빅데이터의 딥러닝 분석을 통한 주가예측 및 수익률 평가

July 6, 2018

홍영현, 양우령, 최성희, 김제혁
(한국산업기술대학교)

이동현 (한국산업기술대학교)

CONTENTS

연구의 필요성

① **재테크**에 대한 관심의 증대

② **'주식의 가치 평가 방법'**의 변화

애널리스트의 타겟 컴퍼니 선정

↓

타겟컴퍼니가 발생시킬 Earning을 통해 자산의 가치 추정

↓

부채의 가치를 차감한 순수 Equity 가치를 사용

→

거대 Marketmaker들의 **인공지능 & 빅데이터 분석** 활용 비중이 증가

연구 가설

전통계량분석과 AI를 통해 투자한 투자수익률이 market return보다 높을 것이다.

1. 서론

2) 선행연구 리뷰

* 참고문헌

1. R 신경망을 이용한 주가예측 모의실험, 2013
(우미영, 박소영, 한영우, 박우창)

요약 : 한국 증시에 상장되어 있는 1000여개의 종목 중,
대표적인 5개 기업(SK, LG, CJ, KT, 삼성 전자)의 과거 주가를
분기별로 조사하고 예측 결과를 실제 값들과 비교한 결과를 제시

데이터

- 예측을 위한 데이터 : 2002년 3월 ~ 2012년 9월까지의 분기별 주가의 평균
- 예측을 위해 입력한 데이터 : 2002년 3월부터 2012년 6월까지
3개월 간격으로 분기마다 평균값을 사용
- 데이터 변수 : 미시적 요인 4개 (PER, ROE, GRS, EPS) ,
거시적 요인 2개 (다우지수, 코스피 지수)

실증 분석

- 분석 방법론 : R에서 제공하는 신경망 예측 모델 nnet 패키지를 호출하여
모델을 훈련 시킨 후, predict를 호출하여 예측

분석 결과 : 2012년 9월의 주가지수를 출력노드가 상승, 하락하는 경우의
예측 정확도가 0.5정도로 나타남

1. 서론

2) 선행연구 리뷰

* 참고문헌

2. 인공지능시스템을 이용한 주가예측에 대한 연구, 2008 (김광용,이경락)

요약 : 인공신경망을 이용한 주식가치 평가와 회계정보를 이용하여
기업가치를 평가하는 Ohlson모형을 적용하여 주식가치를 평가하여
어느 방법이 유용한지 판단

데이터

- 독립변수 : 2001년부터 2004년까지의 ROE 와 매출총이익률 등
각종 기업의 활동성 , 수익성 지표
- 종속변수 : 2001년부터 2004년까지의 각각의 주가

실증 분석

- 분석 방법론 (인공신경망) : 각 년도마다 서로 다른 재무비율의 변수를
이용하고 각각의 인공신경망에 사용된 은닉층도 서로 다르게 사용,
출력변수는 예측하는 종속변수 모두 1로 동일
- 분석 방법론 (Ohlson) : 당기순이익에서 주주의 자본비용을 차감하여
잔여이익을 구하고 기업의 미래가치가 미래 잔여이익과 선형관계라 가정

분석 결과 : 2002년과 2003년, 2004년의 실제주가를 인공신경망과
Ohlson를 통해 예측한 추정주가와 비교해 보았을 때 Ohlson의
모형보다 인공신경망으로 예측한 가격이 실제 주가를 잘 반영함.

2. 방법론

주가 변화에 영향을 미치는
변수를 상관분석과
회귀분석을 통해 파악

독립변수 : 해당 변수와
전분기의 가격
종속변수 : 다음분기의 가격

▶ 로지스틱 회귀분석 및
MLP를 통해
일반식을 찾아냄

해당 모델을 이용하여
과거에 주식을 살 경우와
Market return을 비교

2. 방법론

1) 데이터

1704개 → 1255개

코스피와 코스닥에 상장되어있는 기업

Ex)

종속 변수

9월 대비 12월

주가 상승여부

독립 변수

+ EPS(6월)
+ BPS(6월)
+ Net Income(6월)
+ CFO(6월)
+ 영업이익(6월)
+ 매출액(6월)
+ CAPEX(6월)
+ FCF(6월)
+9월 주식 평균 종가

2. 방법론

1) 데이터

<표 1> 독립변수와 종속변수의 설정

변수	변수의 내용
x1	9월 대비 12월 주가 상승여부
x2	9월의 주식 평균 가격
x3	EPS
x4	BPS
x5	CAPAX
x6	CFO
x7	EBIT
x8	SALES
x9	FCF
x10	NI

2. 방법론

1) 데이터

〈 크롤링을 이용한 재무데이터 수집 〉

```
write.csv(n,"nn.csv")



fndata1<-data.frame()



setwd("c:/kpu_data")
file <-read.csv("대박4.csv")
fndata1<-data.frame()
for(i in 1:1){
  a<-file[i,2]
  url_base<-"http://companyinfo.stock.naver.com/v1/company/ajax/cF1001.aspx?cmp_cd=000"
  url_base2<-"&fin_typ=4&freq_typ=Q"
  setwd("C:/kpu_data")
  url<-paste(url_base,a,sep='')
  url1<-paste(url,url_base2,sep='')
  a<-readHTMLTable(url1,encoding = "UTF-8",header = T)%>%data.frame()
  print(i)
  j=as.data.frame(a)
  n<-j[26,4:3]
  n<-data.frame(n,j[28,4:3])
  n<-data.frame(n,j[17,4:3])
  n<-data.frame(n,j[14,4:3])
  n<-data.frame(n,j[2,4:3])
  n<-data.frame(n,j[1,4:3])
  n<-data.frame(n,j[18,4:3])
  n<-data.frame(n,j[5,5:4])
  colnames(n)<-c("epst","epst-1","bpst","bpst-1","capaxt","capext-1","영업활동현금흐름t",
  fndata1<-bind_rows(fndata1,n)
}

fndata2<-data.frame(fndata1)
```

2017년 6월과 2017년 9월의 재무데이터를
네이버 금융을 통해 자동적으로 수집

2. 방법론

2) 분석방법론

1255개의 Data = 1000개의 훈련데이터 + 255개의 시험데이터

① Python을 통한 MLP Binary Classification 예측
② R을 통한 로지스틱 회귀분석 예측

〈표 2〉 MLP Binary Classification에 활용된 인공신경망 모델의 요약

Input layer node	9
Hidden layer node	2
Output layer node	1
Learning algorithm	gradient descent
Transfer function	sigmoid
Optimizer	adam

3. 실증분석

1) Python을 통한 MLP Binary Classification 예측

〈표 3〉 3가지 방법의 Binary Classification을 통한 수익률 비교

Dependent Variable	Independent Variable	Market return	A.I return
9월 대비 12월 주가 상승 여부	3월의 earning과 9월 주식 평균 종가	4.98%	8.7%
〃	6월의 earning과 9월 주식 평균 종가	4.98%	9.15%
〃	(6월-3월)의 earning과 9월 주식 평균 종가	4.98%	7.46%

3. 실증분석

2) R을 통한 로지스틱 회귀분석 예측

〈표 4〉 3가지 방법의 로지스틱 회귀분석을 통한 주가 예측률 비교

Dependent Variable	Independent Variable	Market return	Quantative Method return
9월 대비 12월 주가 상승 여부	3월의 earning과 9월 주식 평균 종가	4.98%	9.55%
〃	6월의 earning과 9월 주식 평균 종가	4.98%	9.3%
〃	(6월-3월)의 earning과 9월 주식 평균 종가	4.98%	9.55%

3. 실증분석

3) Confusion Matrix를 통한 로지스틱 회귀 모델 예측력 비교

〈표 5〉 Confusion Matrix을 적용한 지표

지표	지표의 의미
accurancy	정확도를 의미하며 전체 데이터 중에서 올바르게 예측한 것이 몇 개 인지를 의미한다.
recall	Recall(=sensitivity)은 실제값이 1인 것 중에 1이라고 예측한 것을 말한다.
precision	1이라고 예측한 것 중에 실제 값이 1인 것을 말한다.
fpr	실제 False데이터를 Positive값으로 예측한 비율을 의미한다.
f1score	F1score는 Recall과 Precision의 가중조화평균을 말하며 F1score값으로 모델의 예측력을 평가할 수 있다.

3. 실증분석

3) Confusion Matrix를 통한 로지스틱 회귀 모델 예측력 비교

〈표 6〉 Confusion Matrix 지표로 나타낸 로지스틱 회귀분석 예측력 비교

Independent Variable	accuracy	recall	precision	fpr	f1score
3월 재무데이터와 9월 주식 평균 가격	53.7%	83.2%	53.8%	83.1%	0.653
6월 재무데이터와 9월 주식 평균 가격	52.9%	88.1%	53.1%	87.5%	0.662
6월,3월의 재무데이터 차이와 9월 주식 평균 가격	53.7%	83.2%	53.8%	83.1%	0.653

3. 실증분석

3) Confusion Matrix를 통한 MLP Binary Classification 예측력 비교

〈표 7〉 Confusion Matrix 지표로 나타낸 MLP Binary Classification 예측력 비교

Independent Variable	accuracy	recall	precision	fpr	f1score
3월 재무데이터와 9월 주식 평균 가격	56.1%	87.4%	54.8%	92%	0.674
6월 재무데이터와 9월 주식 평균 가격	56.9%	65.5%	57.2%	64.5%	0.611
6월,3월의 재무데이터 차이와 9월 주식 평균 가격	53.3%	94.9%	53.1%	96.6%	0.681

4. 결론 및 시사점

- ✓ 분석 결과, 로지스틱 회귀분석과 MLP Binary Classification 중에 인공신경망을 활용한 모델인 MLP Binary Classification이 F1score에서 로지스틱 회귀분석보다 나은 예측력을 보여주었다.

- ✓ 두 모델은 실제 떨어질 기업의 주식을 오를 것이라고 예측한 FPR의 값이 높게 나오고 예측을 성공적으로 한 지표인 Accuracy의 값이 50% 초중반에 머물러 있기 때문에 주가 예측에 적합한 모델은 아니라고 할 수 있다.

4. 결론 및 시사점

- ✓ **로지스틱 회귀분석 모델과 MLP Binary Classification 모델이 예측한 것을 토대로 주식을 구입하였을 때, 예측평균수익률이 전체 시장평균수익률보다 높았다는 점은 향후 이 모델의 적합한 독립변수와 적절한 노드, 활성화함수, optimizer등을 찾았을 때 모델의 예측력 및 수익률이 보다 향상될 것으로 예상한다.**

4. 결론 및 시사점

정확도가 높은 수익률을 보장 하진 않는다.

정확도를 높게 맞추었다 하더라도 큰 폭으로 떨어진 주식을 오른다고 예측했을 경우 전체 수익률은 떨어짐

즉 모델은 방향성만 고려하고 변동성은 고려하지 못하기 때문이다.

변동성을 고려하기란 힘듦 option가격을 이용해야 하고 데이터도 얻기 힘들다.

따라서 앞으로도 계속 방향성을 가지고 연구를 할 필요가 있다.

4. 결론 및 시사점

연구명	종속변수	독립변수	예측력 (등락 기준)	예측력 (주가기준)
R 신경망을 이용한 주가예측 모의실험, 2013	주가의 등락	PER, ROE, GRS, EPS , 다우지수, 코스피 지수	약 50%	대상 아님
인공지능시스템을 이용한 주가예측에 대한 연구, 2008	주식의 가격	매출총이익률, 영업이익률, 재고자산 회전률 등 기업들의 각종 수익성, 활동성 지표	대상 아님	인공신경망이 예측한 주가가 잔여이익 모델에 비해 주가를 잘 예측함
본 연구	주가의 등락	FCF, CFO 등 주주에게 바로 귀속되는 현금흐름과 earning의 지표	52.9% ~ 56.9% (모델, 변수별로 상이)	대상 아님

감사합니다